DOCUMENT RESUME

ED 382 654 TM 023 088

AUTHOR Donoghue, John R.
TITLE Variable Screening for Cluster Analysis.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-94-36
PUB DATE May 94
NOTE 71p.; Portions of the paper presented at the Annual Meeting of the Psychometric Society (Berkeley, CA, June 1993).
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Cluster Analysis; Monte Carlo Methods
IDENTIFIERS Kurtosis; *Screening Procedures; *Variables; Weighting (Statistical)

ABSTRACT

Inclusion of irrelevant variables in a cluster analysis adversely affects subgroup recovery. This paper examines using moment-based statistics to screen variables; only variables that pass the screening are then used in clustering. Normal mixtures are analytically shown often to possess negative kurtosis. Two related measures, "m" and coefficient of bimodality "b," are also examined. A Monte Carlo study compared the screening measures to no selection, De Soete's (1988) ultrametric weights, and the forward selection procedure of Fowlkes, Gnanadesikan, and Kettenring (1988). Screening based on kurtosis degraded recovery and is not recommended. In contrast, screening on "m" or on "b" improved recovery over both no selection and forward selection, and screening performed as well as ultrametric weights. Combining screening with ultrametric weights performed extremely well. All methods were found to be somewhat sensitive to other types of error. Screening variables appears to be a viable alternative to both ultrametric weights and forward selection. The potential advantages and disadvantaged of screening are considered. Eleven tables and six figures illustrate the analyses. An appendix provides supplemental detail. (Contains 51 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RR-94-36

ED 382 654

RESEARCH REPORT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. Coley

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC).

VARIABLE SCREENING FOR CLUSTER ANALYSIS

John R. Donoghue

Educational Testing Service
Princeton, New Jersey
May 1994

BEST COPY AVAILABLE

TM023088

Variable Screening for Cluster Analysis

John R. Donoghue

Educational Testing Service

May 24, 1994

A portion of this paper was presented at the annual meeting of the Psychometric Society in Berkeley, CA, June 1993. Correspondence should be addressed to John R. Donoghue, Mail Stop 02-T, Educational Testing Service, Princeton, NJ 08541. The author was supported by an All-University Fellowship from U.S.C. Part of this work was completed while the author was a predoctoral fellow at Educational Testing Service. The author would like to thank James Carlson, Norman Cliff, Paul Holland, and Frank Jenkins for their many helpful suggestions.

Copyright © 1994, Educational Testing Service. All rights reserved.

Abstract

Inclusion of irrelevant variables in a cluster analysis adversely affects subgroup recovery. This paper examines using moment-based statistics to screen variables; only variables which pass the screening are then used in clustering. Normal mixtures are analytically shown often to possess negative kurtosis. Two related measures, m and coefficient of bimodality b, are also examined.

A Monte Carlo study compared the screening measures to no selection, De Soete's (1988) ultrametric weights, and Fowlkes, Gnanadesikan, and Kettenring's (1988) forward selection procedure. Screening based on kurtosis degraded recovery and is not recommended. In contrast, screening on m or on b improved recovery over both no selection and forward selection, and screening performed as well as ultrametric weights. Combining screening with ultrametric weights performed extremely well. All methods were found to be somewhat sensitive to other types of error.

Screening variables appears a viable alternative to both ultrametric weights and forward selection. The potential advantages and disadvantages of screening are considered.

Keywords: Variable selection; Cluster analysis of two-mode data; Kurtosis; Hierarchical clustering; Euclidean distances.

1. INTRODUCTION

Applications of cluster analysis commonly involve trying to isolate relatively homogeneous subgroups of individuals from a collection of entities hitherto thought to be homogeneous. Thus, this paper adopts the view of cluster analysis as the attempt to "unmix a mixture of distributions" (e.g., Titterington, Smith, & Makov, 1985; McLachlan & Basford, 1988). Clusters are the homogeneous distributions which are mixed, and applications of cluster analysis attempt to identify relatively homogeneous subgroups within a more heterogeneous population.

The first step in such an analysis is to select the necessary entities and variables. Meehl (1979) emphasized the use of clinical insight into the domain of interest, and standard sources on cluster analysis such as Everitt (1980), Lorr (1983), and Aldenderfer and Blashfield (1984) merely state that the variables should be theoretically relevant. Yet, cluster analysis is useful as an exploratory technique; the domain of interest may be known, but the specific variables which separate putative subgroups are not known prior to the analysis.

1.1 The Problem of Irrelevant Variables

The usual response of applied researchers is to include all possible variables, in the hope that the dimensions upon which subgroups differ will be represented by one or more of these variables. Unfortunately, such a shotgun strategy is counter-productive. In the process of clustering, the two-mode (variables by entities) multivariate data are converted to a single-mode (entities by entities) univariate similarity measure, such as Euclidean distance or Q-correlation. Including irrelevant variables acts to introduce noise into the similarity measure, obscuring subgroup structure. Everitt (1980) reports that algorithms such as single and centroid linkage produced similar results when used with similarity data containing error as they did when used to cluster unimodal data. This renders such methods effectively useless, as it is impossible to

interpret such a solution (Donoghue, 1987). Milligan (1980) found that the addition of irrelevant dimensions resulted in the lowest recovery rates for *all* of the clustering algorithms he studied. He concludes that "a researcher should be particularly cautious when selecting variables to be used in the clustering process." (1980, p. 341).

The deleterious effect of irrelevant variables is aggravated by attempts to deal with other problems. Fleiss and Zubin (1969) noted that standardization of variables (to remove effects of variable scale) has the effect of decreasing between-groups spread compared to those variables which do not contain subgroups. This implicitly assigns larger weights to variables which do *not* measure the between-groups difference, making the subgroups harder to isolate. Simulations by Milligan and Cooper (1988) and Barton (1993) have found that standardizing variables can adversely affect recovery by cluster methods. Attempts to deal with problems caused by computing Euclidean distances from non-orthogonal variables (e.g., Donoghue, 1993) produce similar problems. Hartigan (1975) reports decreased recovery when using Mahalanobis distance, and Rohlf (1970) and Chang (1983) discuss problems in clustering based upon principal components scores. However, techniques developed by Art, Gnanadesikan, and Kettenring (1982) to estimate the pooled within-groups covariance matrix may alleviate this problem (Donoghue, 1994). In addition, clustering procedures recently have been proposed which combine multidimensional scaling and/or variable weighting with specific clustering algorithms (De Soete, DeSarbo, & Carroll, 1985; DeSarbo, Carroll, Clark, & Green, 1984; DeSarbo, Howard, & Jedidi, 1991).

1.2 Methods to Deal with Irrelevant Variables

A few general suggestions have appeared which address the problem of irrelevant variables. Unlike those just cited, these methods are not tied to the clustering algorithm used,

and so are applicable across a variety of clustering algorithms. Fowlkes, Gnanadesikan, and Kettenring (1988) suggested a forward selection procedure to determine which variables to include in a cluster analysis. At each step, their method selects the variable which maximizes Pillai's trace criterion from MANOVA. For each analysis, expected values of the statistic are obtained by Monte Carlo methods, i.e., 100 draws of n entities from a spherical, p-dimensional normal distribution. Forward selection stops when the increase in the trace statistic is less than the expected value. The method is computationally intensive, with the amount of computation increasing with the square of the number of variables.

Milligan (1989) examined the use of a variable weighting procedure (De Soete, 1986, 1988) to deal with irrelevant variables. The method selects weights such that the distances computed from the weighted variables maximally satisfy the ultrametric inequality:

$$d_{ij} \leq \max(d_{ik}, d_{jk}) \quad .$$

This is equivalent to requiring that all sets of three points lie on an acute isosceles (or equilateral) triangle. Johnson (1967) and Milligan (1979) demonstrated the relationship between the ultrametric inequality and many commonly used hierarchical clustering algorithms. Milligan (1989) found that using the ultrametric weights improved cluster recovery when the data contained one, two, or three irrelevant dimensions. The amount of computation for this method increases with the cube of the number of entities; Milligan reports that the method required too much computation to complete an addition simulation condition in which datasets contained 250 entities.

1.3 Moment-based Variable Screening

Some researches have suggested screening variables based upon the shape of the distribution. For example, Morris et al. (1981, cited in Fletcher & Satz, 1985) have noted that

normally distributed variables are not consistent with the presence of subgroups in the sample, and Fletcher and Satz assert that the distribution of the variables must be skewed (1985, p. 49) in order for the variables to be consistent with the presence of subgroups. While this paper was in preparation, a study by Bajgier and Aggarwal (1991) was published which compared the power of a variety of univariate distributional tests to detect balanced mixtures. Testing for negative kurtosis was the most powerful of the methods they examined.

In this paper, three variable screening strategies are developed and examined. Section 2 examines the meaning of univariate kurtosis, and its relationship to bimodality. Section 3 examines the kurtosis, and two improvements to the kurtosis, the *m*-index and the coefficient of bimodality *b*. Section 4 gives the design of a simulation study to examine these methods. Section 5 reports the results of the simulation, and compares the screening measures to two alternatives from the literature. Finally, Section 6 contains discussion of the potential advantages and disadvantages of screening, and Section 7 presents suggestions for further work.

2. THE DISTRIBUTION OF A MIXTURE

Mixtures are expected to have multiple modes corresponding to the individual subgroups. Finucan (1964, p. 112) noted, "a bimodal curve in general has also a strong negative kurtosis." A series of notes in *The American Statistician* also suggest this (Darlington, 1970; Chissom, 1970; Hildebrand, 1971), but they also point out that kurtosis is not necessarily negative for bimodal distributions. In addition, Eisenberger (1964) has examined the conditions under which a mixture of two normal distributions will be bimodal or unimodal. Distribution B in Figure 1 is such a unimodal mixture of two normal distributions. In 1939, Fisher asserted that distributions such as B had a lower kurtosis than distribution A, the standard normal distribution. Finucan (1964) proved this assertion, as have others from different points of view (Marsaglia, Marshall,

& Proshau, 1965; Ali, 1974). As a result, kurtosis is often interpreted as a measure of whether the distribution is sharply peaked or flattened out compared to the normal distribution. Yet, Kaplansky (1945) demonstrated that the kurtosis need not be related to the distribution's peakedness, and Ali (1974) and Johnson, Lietjan and Beckman (1980) have argued that kurtosis is better conceived of as a measure of the thickness of the tails of the distribution. Distributions which have thicker tails than the normal take on positive values of kurtosis; those with flatter tails take on negative values. Balanda and MacGillivray's (1988) review concluded that "it is best to define kurtosis vaguely as the location- and scale-free movement of probability mass from the shoulders of a distribution into its center and tails, and to recognize that it can be formalized in many ways." (p. 111)

Insert Figure 1 about here

A mixture of normal distributions may be unimodal or bimodal. In some cases unimodal mixtures of normals can have lower kurtosis than a single normal of equal mean and variance. Bimodal distributions generally may have negative kurtosis, although not always. Hence, there appears to be some connection between negative kurtosis and mixtures. Thus, we next consider the kurtosis of a mixture.

3. KURTOSIS OF A MIXTURE

The measure of kurtosis, g_2, is the fourth moment about the mean normalized by the variance squared, and compared to the normal's normalized fourth moment (which is 3):

$$g_2 = \frac{M_4}{M_2^2} - 3 \quad , \qquad (1)$$

where M_k is the kth moment about the mean. The kurtosis of a mixture of normal distributions

is given by:

$$g_{2m} = \frac{3\ var\left[\sigma_j^2 + (\mu_j - \mu)\right] - 2\sum_{j=1}^{G}\pi_j(\mu_j - \mu)^4}{\left[\sum_{j=1}^{G}\pi_j\sigma_j^2 + \sum_{j=1}^{G}\pi_j(\mu_j - \mu)^2\right]^2}, \qquad (2)$$

where *var* is the variance over subgroups and π_j is the proportion in subgroup j (the derivation of (2) is given in the Appendix). There are three competing processes working to determine the kurtosis of a mixture. Heterogeneity of the within-group variances inflates the kurtosis, as do differences in the sizes of the subgroups. Differences in the subgroup means work to decrease the kurtosis.

The form of (2) allows two properties to be easily demonstrated:

A) Common Mean: When the subgroup means are identically equal to μ, all of the $(\mu_j - \mu)^2$ terms drop out, and the kurtosis of the mixture is non-negative, $g_{2m} \geq 0$.

B) Homogeneous Variances: When the subgroup variances are identically equal to σ^2, the kurtosis is a function only of the subgroup means. In this case, g_{2m} will be less than zero, provided that the π_j are not too dissimilar. Thus, the kurtosis will be negative whenever the variance of the squared differences of the means $(\mu_j - \mu)^2$ is less than two–thirds of the sum of $(\mu_j - \mu)^4$. While this expression has no simple, intuitive meaning, it will be true whenever the subgroups are relatively similar in size, for example if the ratio of the sizes is less than 3:1 in the two subgroup case.

Assuming normality of within–group distributions, platykurtosis ($g_{2m} < 0$) indicates the presence of subgroups. Unfortunately, the converse is not true. Thus, kurtosis may be used as a relatively stringent screening measure. The inferences which may be made (in the absence of

sampling error) are summarized in Table 1.

Insert Table 1 about here

The screening test based upon kurtosis may be improved by including the information about the variable's skewness. Such a correction is particularly attractive because kurtosis is most powerful in situations in which the overall distribution is nearly symmetric (in Table 2, cell III and cell II when the π_j are similar). The skewness, g_1 is defined as:

$$g_1 = \frac{M_3}{[M_2]^{3/2}} \quad .$$

For a mixture, this becomes:

$$g_{1m} = \frac{3\sum_{j=1}^{G} \pi_j \sigma_j^2 (\mu_j - \mu) + \sum_{j=1}^{G} \pi_j (\mu_j - \mu)^3}{\left[\sum_{j=1}^{G} \pi_j \sigma_j^2 + \sum_{j=1}^{G} \pi_j (\mu_j - \mu)^2\right]^{3/2}} \quad .$$

The factors of unequal within–group variances and unequal mixing proportions (π_j) induce skewness in the mixture distribution. Thus, incorporating a correction for the skew will make the test more powerful.

The kurtosis is always bounded below. This lower bound is usually given as:

$$g_2 \geq -2.$$

However, the actual lower bound (Stuart & Ord, 1987, p. 115) is:

$$g_2 + 3 \geq g_1^2 + 1 \, . \tag{3}$$

This suggests the index m:

$$m = g_2 - g_1^2 \, , \tag{4}$$

which has a uniform lower bound of -2.0 for all variables.

As with kurtosis, a plausible selection rule is to use variables for which $m \leq 0$. There is no simple expression for (4) in the presence of mixtures. However, it does have several desirable properties. It has an expected value of zero for a single (nonmixture) normal distribution. Under the condition of homogeneity of subgroup means, it reduces to the kurtosis of the mixture, and again has a nonnegative expected value. Similarly, under homogeneity of subgroup variances, the value is negative for most cases.

The coefficient of bimodality b in SAS (1985) also incorporates the bound in (3):

$$b = \frac{g_1^2 + 1}{g_2 + 3} \quad .$$

The coefficient is bounded, $0 < b \leq 1$. The manual suggests that values "greater than 0.555 may indicate bimodal or multimodal" distributions (1985, p. 272). No explanation is given for this value, but it is the expected value of the statistic for a uniform distribution, and assumes that values larger than this are likely to reflect true subgroup structure (W. S. Sarle, personal communication, June 11, 1987). To date, no studies have investigated the efficacy or power of this measure. The expected value of the statistic is .333 for a single normal distribution. Large values of b suggest multimodality. It will take on values less than or equal to .333 when no mean differences are present. Also, b is more sensitive than is m; $m < 0$ implies $b > .333$.

To illustrate the behavior of g_2, m, and b, the expected values of each were calculated for a variety of mixtures. Table 2 gives values for each of the measures for selected combinations of subgroup proportions, means and variances for mixtures of two and three subgroups, and Table 3 gives the values of each of the statistics for several common probability distributions.

Insert Tables 2 and 3 about here

Table 3 illustrates a desirable property of the measure m. Its expected value is independent of the distributional parameters for several of the distributions examined. In the remaining cases it is a very simple function. This is not true of b. It is not clear whether this property is more important than the slightly greater sensitivity of b.

4. SIMULATION DESIGN

A Monte Carlo study was undertaken to systematically evaluate the proposed screening measures on the ability of common clustering algorithms to recover a known subgroup structure.

4.1 Method

Data were generated using a modified version of the algorithm given by Milligan (1985). This algorithm has been used in a number of studies. Each dataset consisted of 50 observations. Within a subgroup, observations were drawn from a truncated multinormal distribution, with observations constrained to lie within the range, $\mu_j \pm 1.5\ \sigma_j$ in the first dimension. In addition, subgroup boundaries were well separated in the first dimension. This insured that there was no overlap among the subgroups.

Design

The chief variable of interest was the effect of the variable selection/weighting procedures which were applied to the datasets. Four additional factors were manipulated in the data generation:

1) Number of subgroups (4 levels) -- 2, 3, 4, or 5 subgroups,

2) Number of "core" variables (3 levels) -- 4, 6, or 8. Subgroup means differed on each of these dimensions.

3) Density of subgroups (3 levels):

a) equal sized subgroups,

b) the first subgroup was 10% of observations, other subgroups were equal sized,

c) the first subgroup was 60% of observations, other subgroups were equal sized.

These three factors were fully crossed to yield 36 (4 X 3 X 3) conditions. Three replicate datasets were generated per condition, for a total of 108 base datasets. Each of the base datasets was then modified according to seven error conditions:

4) Error condition

a) No error

b) One normally distributed noise dimension

c) Two normally distributed noise dimensions

d) Three normally distributed noise dimensions

e) Error perturbed coordinates, low error, $\lambda = 1$

f) Error perturbed coordinates, high error, $\lambda = 2$

g) Outlier condition, 10 observations (i.e., 20%) which did not fall within any subgroup were added to the dataset.

For the error perturbed coordinates condition, the normally distributed error was added to the original coordinates:

$$E_{jik} = A_{jik} + \lambda e_{jik}$$

$$e_{jik} \sim N(0, \sigma^2_{jk})$$

This resulted in a total of 756 datasets. See Milligan (1985) for additional details on the data generation and error conditions. The variables in each dataset were then weighted and/or

screened using each of the 10 methods listed below, yielding 7,560 weighted datasets. Each weighted dataset was then analyzed by four clustering algorithms, making a total of 30,240 clusterings. For each clustering, the solution for the correct number of subgroups was used as the result for that method.

Selection Algorithms

Each data set was subjected to 10 variable weighting/selection strategies:

A) No selection,

B) $g_2 < -1.2$,

C) g_2 significantly < 0. This was determined at $\alpha = .05$, using the tables in Chen (1983),

D) $g_2 < 0$,

E) $m < -1.2$,

F) $m < 0$,

G) $b > .555$,

H) $b > .333$,

I) De Soete's (1986, 1988) ultrametric weighting algorithm,

J) Fowlkes, Gnanadesikan, and Kettenring's (1988) forward selection algorithm.

Cluster Algorithms

The 10 weighted versions of each dataset were then analyzed four times, corresponding to different hierarchical clustering algorithms and measures of similarity. The clustering methods were: (a) Single linkage, Euclidean distance; (b) Complete linkage, Euclidean distance; (c) Average linkage, Euclidean distance; (d) Ward's method (minimum variance), squared Euclidean distance. The clustering methods were chosen because they are widely used and average linkage and Ward's method have consistently performed well in previous studies. For a

discussion of these algorithms, the reader is referred to standard introductions to cluster analysis (e.g., Everitt, 1980; Lorr, 1983).

Outcome Measure

The outcome measure for the study was the Hubert and Arabie (1985) modification of Rand's (1971) statistic, which will be denoted HA-Rand. The index was computed between each cluster solution and the true subgroup membership used to generate the data. This index is based on examining pairs of entities, and determining whether they are classified into the same or different subgroups. A value of zero reflects chance agreement with the true membership, and 1.0 reflects perfect agreement. A study by Milligan and Cooper (1986) supports the accuracy of Hubert and Arabie's modification.

Computer Programs

Data were generated using a modified version of Milligan's (1985) program. The weights for De Soete's algorithm were computed using his program OVWTRE (De Soete, 1988).[1] The moment statistics, the Fowlkes, Gnanadesikan, and Kettenring (1988) forward selection algorithm, and clustering algorithms were computed using FORTRAN programs written by the author. Accuracy of these programs was ensured through numerous comparisons of results of subroutines and final classifications with routines from SAS and SPLUS. Eigenvalues were computed using routines from EISPACK (Smith, Boyle, Garbow, Ikebe, Klema, & Moler, 1974). Note that the forward selection procedure in Fowlkes, Gnanadesikan, and Kettenring (1988) was developed in terms of the complete link clustering method. For the present study, the full method of determining expected values via Monte Carlo methods and

[1] The author is indebted to Glenn Milligan for providing a copy of the source code of his generation program and OVWTRE.

then performing forward selection was applied to each of the clustering algorithms.

5. SIMULATION RESULTS

Although ANOVA might seem a natural means to summarize the results, it was not used for the present study.[2] The primary independent variable of interest was the weighting/selection method used in the analysis. The other factors in the design were included to ensure that their effects were systematically present, and so would not confound the results concerning variable weighting/selection. It is more meaningful to directly examine the comparisons of interest, using multiple comparison procedures to control overall Type I error rate. Still, there may be interest in the main effects of the other variables in the study. These are summarized in Appendix Table A1. In general, the effects replicate those in other studies (e.g., Milligan, 1980, 1989).

The variable screening/weighting methods primarily were compared using a distribution free ordinal procedure, Cliff's (1993) method of comparing the order of two distributions. Ordinal comparisons were performed using a modified version of Cliff's (1992) program PAIRDEL1, for paired observations. Two types of ordinal hypotheses were assessed. The first, based on the index d_w, is the proportion of datasets for which one method yielded higher recovery than the other method minus the proportion for which it yielded lower recovery; it is the net proportion of datasets with improved recovery. Negative values of d_w indicate lower recovery for first method. The second ordinal procedure estimates the probability that a randomly sampled observation from one distribution has a larger value than a randomly sampled

[2] In addition, a preliminary investigation of the within-cell means revealed substantial heterogeneity of variance, violating the ANOVA assumption, and making the ANOVA tests suspect.

observation from the other distribution. This results in one of three decisions for each pair of clustering methods: a) Method A is higher (better recovery) than Method B; b) Method B is higher than Method A; or c) the methods do not significantly differ. In addition, pairwise t-tests of means were also computed.

The results will be discussed in six sections. In Section 5.1, the minimal requirement of effectiveness for the proposed screening measures is examined: Does using the measure yield an improvement over no screening? Measures which provide no improvement are certainly not worth adopting. Next, Section 5.2 compares recovery using the variable screening to two suggestions from the literature: (a) variable weights to maximize agreement of the distances with the ultrametric inequality (De Soete's 1988 program); and (b) the forward selection procedure of Fowlkes, Gnanadesikan, and Kettenring (1988). Section 5.3 looks at the robustness of the screening procedures; how do they perform in the presence of other types of error (perturbed coordinates and outliers)? Section 5.4 evaluates the effect of combining variable screening with ultrametric weights. Next, Section 5.5 examines the interaction of the best screening/weighting methods with clustering algorithms. Finally, Section 5.6 explores the effect of variable standardization on the behavior of the forward selection algorithm.

5.1 Effectiveness

The minimal requirement of effectiveness is that using the variable screening/weighting method yield an improvement over using no selection. To address this issue, each method was compared to no selection. These comparisons were made on HA-Rand index values pooled over all datasets containing 0, 1, 2, or 3 error dimensions. Table 4 summarizes the results of these comparisons.

Insert Table 4 about here

The three methods based on the kurtosis, $g_2 < -1.2$, g_2 signif. < 0, and $g_2 < 0$, yielded significantly worse recovery than no selection. Negative d_w values indicate that the kurtosis-based measures resulted in lower HA-Rand values for 9-27% of the datasets. This result differs sharply from the results of Bajgier and Aggarwal (1991), who found kurtosis to be the most powerful measure for detecting mixtures. However, Bajgier and Aggarwal only examined balanced mixtures, i.e., mixtures with equal mixing proportions and equal variances. To determine whether this accounted for the difference in findings, Figure 2 plots mean HA-Rand index results for no selection and for each of the kurtosis-based measures by subgroup size. Consistent with Bajgier and Aggarwal, the kurtosis-based measures function well for equal-sized subgroups. When the subgroups differ in size, however, these procedures do not function very well. Overall, therefore, kurtosis-based measures do not meet the basic test of effectiveness as screening procedures and will not be discussed further.

Insert Figure 2 about here

As was noted above, unequal subgroup sizes induce skewness in the overall distributions. Thus, the measures which incorporate information about skewness, b and m, may be more useful. Table 4 reveals that all four of the screening methods involving b or m yield better recovery than no selection. In addition, neither the m-index nor b showed a large effect for subgroup size. The largest effect for subgroup size was a difference in HA-Rand index of approximately .06; for the kurtosis-based measures the effects ranged from .15 to .30.

Weighting the variables to maximize agreement with the ultrametric inequality yielded

significantly higher recovery than no selection. The forward selection method did not differ from no selection in the ordinal comparisons. However, paired t-tests indicated that forward selection did yield a significantly higher mean HA-Rand index than no selection. The differences in these results will be examined in more detail in the next section.

5.2 Comparison with Other Weighting/Selection Methods

Pairwise comparisons of the 7 remaining methods[3] were made on HA-Rand index values for analyses of all datasets containing 0, 1, 2, or 3 error dimensions. Shaffer's (1986) modification to the Bonferroni correction was used to maintain familywise Type 1 error rate of $\alpha = 0.05$. Finally, these pairwise relations were converted into ranks, based upon the number of methods which were significantly higher than a given method versus the number of methods which were significantly lower. These results are summarized in Table 5.

Insert Table 5 about here

Table 5 also presents mean recovery for each number of error dimensions. When there are no error dimensions, $m < 0$ yields similar recovery to no selection, while $m < -1.2$ gives somewhat worse recovery. Screening based on the test of normality, $m < 0$, recovery is somewhat affected by the addition of error dimensions, but less so than no selection. On the other hand, screening based on the uniform distribution ($m < -1.2$) yields similar results for all numbers of error dimensions. Overall, in the presence of error dimensions, however, both the normal and uniform tests yield HA-Rand recovery values higher than those for no selection. The pattern of results for b, the coefficient of bimodality, is very similar to that for the m-index,

[3] Kurtosis-based methods are not discussed due to their poor performance in the comparison with no selection.

although the normality test (b > .333) provides only minimal improvement over no selection.

In comparing the relative effectiveness of the screening measures with the ultrametric weights and forward selection procedures, Table 5 reveals that recovery using the ultrametric weights did not significantly differ from the screening methods based on a uniform distribution (b > .555 and m < -1.2), but outperformed both methods based on a normal distribution (b > .333 and m < 0). The paired t-test results indicated that screening based on m < 0 did not differ from the ultrametric weights, but screening based on b > .333 was still worse. Based on the ordinal comparisons, the forward selection method was found to yield lower cluster recovery than all four of the variable selection methods using b and m. However, based on the paired t-test results, the forward selection method is superior to selection based on b > .333, and did not differ from the other methods.

A word is in order concerning discrepancies between the rank orders derived from the ordinal comparisons and those implied by paired t-test of the means. Forward selection has a noticeably higher mean than selection based on b > .333, yet the ordinal comparison indicates that recovery for forward selection is significantly *lower*. This seeming paradox points out the different questions addressed by the two comparisons. The ordinal method compares differences in direction, but the means take into account the size of those differences. Examination of the differences in the individual solutions confirms that b > .333 yields more cluster solutions with HA-Rand that is higher than forward selection than vice versa. However, forward selection occasionally produces a solution which is much better, giving forward selection a higher mean. Thus, both are legitimate answers to the question: Which method is better?

5.3 Robustness to Other Types of Error

An additional issue in comparing the methods is their sensitivity to other types of error

contaminating the cluster structure. Milligan's cluster generating program includes three additional error conditions: (a) data perturbed by adding an error to each coordinate, low error variance; (b) data perturbed by adding an error to each coordinate, high error variance; and (c) including an additional 20% (i.e., 10 cases) which do not lie within the boundaries of any of the clusters (i.e., outliers and intermediates). These conditions will be referred to as, respectively, low error, high error, and outlier conditions.

Table 6 gives the mean HA-Rand index and ranks based on ordinal pairwise comparisons of the methods for each of the error conditions. For all conditions, $b > .555$ and $m < -1.2$ yield much lower recovery than other methods, indicating that these selection methods degrade cluster recovery in the presence of error other than spurious variables. On the other hand, variable selection methods based on normality, $b > .333$ and $m < 0$, are relatively robust, and show little difference in cluster recovery from that of no selection.

Insert Table 6 about here

5.4 Combined Methods

Variable selection based on the m-index and the coefficient of bimodality b are effective in reducing the effects of spurious dimensions. Variable weighting based on the ultrametric inequality is also effective. This section examines the effectiveness of combining the two strategies, selection and ultrametric weights. Variations of the forward selection method were not considered; forward selection is extremely computationally intensive, and combining the method with other techniques was not feasible for the purposes of this study.

Each of the datasets was reanalyzed. First, one of four variable screening methods ($m < -1.2$, $m < 0$, $b > .333$, or $b > .555$) was applied. The variables passing the screening were

then analyzed using De Soete's (1988) program to determine the variable weights. If no variables passed the screening, all variables were used. The weighted distances were then computed and analyzed by the clustering algorithms, as described in the Method section.

In order for the combination of weighting and screening to be effective, the results of the combined methods must be superior to both weighting alone and screening alone. Table 7 summarizes comparisons of the combined methods to (a) screening alone and (b) using only the ultrametric weights. Although most of the methods show improvement, the combination of weights and screening based on $m < -1.2$ yielded *worse* recovery than did either method alone. The negative values of d_w and the ordinal z-test indicate lower recovery for the combined method, although the paired t-test indicates that the combined method yields a higher mean than does $m < -1.2$ alone. This pattern of results suggests that the combined method often yields somewhat lower recovery, but occasionally does much better than screening alone. Combining weighting with screening based on $b > .555$ yielded increased recovery in a net 2-3% of the datasets, and gives higher mean recovery, although the overall comparison of distributions does not significantly differ from screening alone. Finally, combining weighting with either of the two methods based on a normal distribution ($m < 0$ and $b > .333$) clearly improves recovery.

Insert Table 7 about here

Table 8 summarizes the results of applying the combined procedures to datasets with 0, 1, 2, or 3 error dimensions. In addition, results for five additional methods (No Selection, ultrametric weights, forward selection, and the unweighted versions of $m < -1.2$, and $b > .555$) are repeated from Table 5. Overall, best recovery was obtained for $m < 0$ with weights and

$b > .555$ with weights. Although the ordinal test indicated that the latter did not differ from the unweighted version, the mean for the combination method is much higher. The profile of means for $b > .555$ is impressive; there is virtually no effect of increasing from 0 to 3 error dimensions. On the other hand, $m < 0$ shows a modest effect of increasing error dimensions.

Insert Table 8 about here

Table 9 presents the results for the other error conditions: low error, high error, and 20% outlier. The combined method based on $b > .555$ shows considerable sensitivity to the other types of error, and is uniformly among the three methods with the lowest recovery. The combined method based on $m < 0$ is much less sensitive to the other types of error, and does not differ from ultrametric weights only for the low error or high error conditions. Comparison with Table 6 reveals that the means are very similar to the unweighted version for these two conditions, although the combined method does appear to be somewhat affected by the presence of outliers.

Insert Table 9 about here

5.5 Interaction of Variable Screening with Clustering Methods

An additional question of interest is whether the variable weighting/screening methods differed in usefulness for the different clustering algorithms. To address this issue, the mean for each clustering algorithm was computed for six of the eight methods listed in Table 6. Screening based on $b > .555$ and $m < -1.2$ were omitted. For each of the clustering algorithms, the omitted methods showed a very similar pattern to other screening methods, which also yielded higher recovery. Means for the average linkage algorithm are plotted in Figure 3. Results for

Ward's method are shown in Figure 4. Recovery for the complete linkage algorithm is portrayed in Figure 5 and means for single linkage are plotted in Figure 6.

Insert Figures 3 through 6 about here

There were few large interactions between clustering algorithms and the variable weighting/screening methods. The notable exception is the behavior of forward selection for the single linkage algorithm. For the other algorithms forward selection appears to have little to recommend it; using forward selection with average linkage yields recovery which is uniformly lower than that of any other method, including no selection. On the other hand, the method gives uniformly high recovery when used with single linkage clustering, and is the best method for that algorithm. Other method by clustering algorithm interactions were relatively small.

5.6 Behavior of the Forward Selection Method

The relatively poor performance of the forward selection method of Fowlkes, Gnanadesikan, and Kettenring (1988) was unexpected. Results presented in their paper indicated that the method was very effective, if somewhat computationally intensive. The datasets in their study tended to have variables with similar variances. In this study, both within-group and overall variances were allowed to differ rather widely. The forward selection procedure standardizes each of the variables by its total variance in order to remove spurious scale effects from the computation of eigenvalues used in the selection. Milligan and Cooper (1988) and Barton (1993) have found that this method of standardization can adversely affect recovery by cluster methods.

Of the methods used in this study, only the forward selection procedure used standardized variables. To investigate whether this difference might have caused the unexpected

performance of the forward selection data, the datasets were reanalyzed. Only those variables selected by the forward selection procedure were included in the analysis, but the variables were *not* standardized in forming the Euclidean distances. Table 10 compares results from this method with the standardized version of forward selection, no selection, ultrametric weights, and the two best combined methods for screening and weights. Forward selection was adversely affected by variable standardization; standardizing variables leads to lower recovery in almost 9% more datasets than vice versa. Comparisons with other methods reveal that without standardizing variables, forward selection is superior to no selection, and ordinal comparisons indicate that it does not differ significantly from the other methods. Mean comparisons indicate marginally better recovery than ultrametric weighting alone and marginally worse recovery than screening based on $b > .555$ combined with weights. Means for each number of irrelevant dimensions are presented in Table 11.

Insert Table 10 and Table 11 about here

---

6. DISCUSSION

Replicating the work of other authors, the inclusion of irrelevant dimensions was found to severely degrade cluster recovery. This paper examined the usefulness of moment-based univariate statistics to screen variables. Only variables which pass the screening are then used in the clustering. Results for screening based on the kurtosis measure g_2 were very poor. For subgroups of equal size, g_2 functioned fairly well, but did very poorly for unequal sized subgroups. Thus, it appears that the results of Bajgier and Aggarwal (1991) do not generalize, and screening based on g_2 cannot be recommended for applied clustering.

In contrast, screening based on the index m and on the coefficient of bimodality b

functioned well. Both measures provided increased recovery over no selection and forward selection, and versions of each (m < -1.2 and b > .555) performed as well as the ultrametric weights. However, there is evidence that all of the weighting/selection methods are sensitive to types of error other than spurious dimensions. Selection based on b > .555 and m < -1.2 were most severely affected, although forward selection and ultrametric weights were also affected. Selection based on b > .333 was least affected, followed by m < 0.

Combining variable screening with ultrametric weights performed very well. Two combinations specifically, m < 0 with weights and b > .555 with weights showed improved cluster recovery in the presence of irrelevant dimensions. However, the combined methods (particularly b > .555 with weights) were sensitive to types of error other than irrelevant dimensions. The combined method based on m < 0 was better, although it does appear to be somewhat more sensitive to outliers than either screening alone or ultrametric weights alone. However, procedures have been developed to identify outliers prior to clustering (e.g., Barton, 1991). The use of such procedures may further improve the performance of the combined methods.

One limitation of the present study is that the overall sample size, excluding outliers, was held constant. It is possible that the various variable selection/weighting methods examined here may be dependent on this aspect of the data. Further work should examine the extent to which this is true.

These results indicate that variable screening based on b and m are viable alternatives to both the ultrametric weighting method and the forward selection method. It is worthwhile to briefly consider the relative advantages and disadvantages of screening, compared to the other methods. The advantages of m and b are ease and speed of computation, ready availability, and

potential applicability to a wide variety of clustering methods. Potential disadvantages include the large sampling variability of higher moments, the dependence on the mixture model of clustering, and the potential of univariate methods to fail to identify variables which, although individually providing little information, as a set yield large subgroup separation.

6.1 Potential Advantages

The measures m and b are based on the moment statistics, the skewness and the kurtosis. Thus, they are simple and quick to compute, and the amount of computation in examining a given dataset increases linearly with the number of entities and with the number of variables. In contrast, both ultrametric weights and forward selection are computationally intensive, making their use problematic for large datasets. Computation for the ultrametric weighting algorithm increases with the cube of the number of entities, while computation for the forward selection method increases with the square of the number of variables. Indeed, well over 95% of the computational effort of the simulation results reported here were devoted to the forward selection method.

The components of m and b, the skewness measure g_1 and the kurtosis, g_2, are widely available as standard descriptive statistics. Thus, these measures may be adopted easily by researchers. The ultrametric weights require alternating two multivariate optimization problems, a task which may well be beyond many applied researchers. The method is not widely available, i.e., in statistical packages, although De Soete (1988) has a program to compute the weights. The forward selection procedure is even harder to implement. Determining the expected value of the trace statistic requires drawing multiple multivariate samples, performing a MANOVA decomposition of the results of clustering each sample, and computing the resultant eigenvalues. This is only moderately demanding in an interactive statistical environment such as S-PLUS or

GAUSS, a major undertaking in FORTRAN or C, and borders on herculean in a statistical package such as SPSS or BMDP.

The measures *m* and *b* are suggested based on analysis of the mixture model of clustering. Thus, use of the measures is justified with other types of clustering algorithms, such as iterative partitioning algorithms (i.e., k-means) or direct application of finite mixture models, although the empirical utility of using the methods in these settings has yet to be established. The ultrametric weights, on the other hand, are closely tied to hierarchical clustering. The proofs by Johnson (1967) and Milligan (1979) specifically relate to hierarchical clustering. There is no logical reason to expect ultrametric weights to improve clustering by nonhierarchical algorithms, although it may be empirically found to be useful. The forward selection method is closely related to the normal mixture conception of clustering. Thus, its application is logically valid, although some operational details of the application of the method would need be to resolved.

6.2 Potential Disadvantages

The relationship of the variable screening measures to mixture models may also be a disadvantage. Ultrametric weighting may apply in other conceptions (e.g., graph-theoretic) of hierarchical clustering. In these cases, the mixture model conception may not make sense. The utility of *m* and *b* would have to be established empirically in such situations. Similarly, some applications of cluster analysis are inherently hierarchical (e.g., evolutionary biology), and again the use of the screening measures would have to be established empirically.

Another potential disadvantage of the screening measures is their dependence on the third and fourth moments about the mean, which are rather poorly estimated in samples. This raises valid concern over the degree to which *m* and *b* may fluctuate simply due to sampling

variation. The simulation results presented in Section 5 offer some encouragement along these lines. Each dataset contained a total sample of 50 entities, yet both measures were successful in screening irrelevant variables. Still, more knowledge about the variability of these screening measures would be helpful.

Finally, a potential disadvantage in using univariate techniques such as m or b or the forward selection procedure is that a linear combination of two or more variables may provide good separation between subgroups, while neither of the marginal distributions reveals much separation.[4] There is a danger that the screening may drop such variables, and so lose information about the subgroup separation. It is unknown how the ultrametric weighting method would be affected by such a combination of variables. The forward selection procedure may be less prone to this type of behavior. It is based on a MANOVA test statistic, and is sensitive to subgroup separation based on linear combinations. However, if neither variable provides sufficient univariate separation for inclusion, the forward selection procedure will not detect that the pair provides good separation. It remains for future research to determine how adversely affected the weighting/selection methods are by such combinations of variables.

---

[4] The author would like to thank an anonymous reviewer for pointing out this possibility.

7. CONCLUSION

Taken as a whole, the results of this work are promising. The screening measures *m* and *b* were successful in alleviating the deleterious effect of including irrelevant variables. Both measures provided increased recovery over no selection and forward selection, and versions of each performed as well as the use of ultrametric weights. The success of the combination variable screening and ultrametric weights commends the use of these combined techniques; the combination yielded better recovery than either method separately in a net 2.4-6.7 percent of the datasets analyzed. However, it is not known to what extent these findings are sensitive to specific aspects of this study. This is especially true of the distribution of the irrelevant variables, and the structure of the subgroup separation. Clearly, more work along these lines is warranted.

References

Aldenderfer, M. S., & Blashfield, R. K. (1984). Cluster analysis. Sage University Paper Series on Quantitative Applications in the Social Sciences 07–044. Beverly Hills: Sage Publications.

Ali, M. M. (1974). Stochastic ordering and kurtosis measure. Journal of the American Statistical Association, 69, 543–545.

Art, D., Gnanadesikan, R., & Kettenring, J. R. (1982). Data based metrics for cluster analysis. Utilitas Mathematica, 21A, 75–99.

Bajgier, S. M., & Aggarwal, L. K. (1991). Powers of goodness-of-fit tests in detecting balanced mixed normal distributions. Educational and Psychological Measurement, 51, 253-269.

Balanda, K. P., & MacGillivray, H. L. (1988). Kurtosis: A critical review. American Statistician, 42, 111-119.

Barton, R. M. (1991, April). Outlier detection in cluster analysis using weighted multidimensional scaling. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Barton, R. M. (1993, April). Standardizing variables in cluster analysis. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Chang, W-C. (1983). On using principal components before separating a mixture of two multivariate normal distributions. Applied Statistics, 32, 267-275.

Chen, W. W. S. (1983). On the comparison of some measures of kurtosis: A test of normality. American Statistical Association 1983 Proceedings of the Statistical Computing Section, pp. 217-222.

Chissom, B. S. (1970). Interpretation of the kurtosis statistic. The American Statistician, 24 (10), 19–22.

Cliff N. (1992). PAIRDEL1.BAS: Program for computing matched-data d-statistics [computer program]. Los Angeles: Psychology Department, University of Southern California.

Cliff, N. (1993). Dominance relations: Ordinal analyses to answer ordinal questions. Psychological Bulletin, 114, 494-509.

Darlington, R. B. (1970). Is kurtosis really "peakedness"? The American Statistician, 24 (2), 19–20.

De Soete, G. (1986). Optimal variable weighting for ultrametric and additive tree clustering. Quality and Quantity, 20, 169-180.

De Soete, G. (1988). Software abstract - OVWTRE: A program for optimal variable weighting for ultrametric and additive tree fitting. Journal of Classification, 5, 101-104.

De Soete, G., DeSarbo, W. S., & Carroll, J. D. (1985). Optimal variable weighting for hierarchical clustering: An alternating least-squares algorithm. Journal of Classification, 2, 173-192.

DeSarbo, W. S., Carroll, J. D., Clark, L. A., & Green, P. E. (1984). Synthesized clustering: A method for amalgamating alternative clustering bases with differential weighting of variables. Psychometrika, 49, 57-78.

DeSarbo, W., Howard, D. J., & Jedidi, K. (1991). MULTICLUS: A new method for simultaneously performing multidimensional scaling and cluster analysis. Psychometrika, 56, 121-136.

Donoghue, J. R. (1987). Cluster analysis of learning disabled children. Unpublished masters thesis. California State University: Northridge, CA.

Donoghue, J. R. (1993, April). A Monte Carlo study of the effects of within-group covariance structure on recovery in cluster analysis. Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Donoghue, J. R. (1994, April). Comparing the effectiveness of cluster analysis weighting procedures for within-group covariance structure: The Bivariate Case. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Eisenberger, I. (1964). Genesis of bimodal distributions. Technometrics, 6, 357–363.

Everitt, B. S. (1980). Cluster analysis, (2nd ed.). London: Halstead Press.

Finucan, H. M. (1964). A note on kurtosis. Journal of the Royal Statistical Society, Series B, 26, 111–112.

Fisher, R. A. (1939). Statistical methods for research workers. London: Oliver and Boyd.

Fleiss, J. L., & Zubin, J. (1969). On the methods and theory of clustering. Multivariate Behavioral Research, 4, 235–250.

Fletcher, J. M., & Satz, P. (1985). Cluster analysis and the search for learning disabilities subtypes. In B. P. Rourke (Ed.), Neuropsychology of learning disabilities: Essentials of subtypes analysis (pp. 40–64). New York: Guilford.

Fowlkes, E. B., Gnanadesikan, R., & Kettenring, J. R. (1988). Variable selection in clustering. Journal of Classification, 5, 205-228.

Hartigan, J. A. (1975). Clustering algorithms. New York: John Wiley and Sons.

Hildebrand, D. K. (1971). Kurtosis measures bimodality? The American Statistician, 25 (2), 42–43.

Hubert, L., & Arabie, P. (1985). Comparing partitions. Journal of Classification, 2, 193–218.

Johnson, S. C. (1967). Hierarchical clustering schemes. Psychometrika, 32 241-254.

Johnson, M. E., Lietjan, G. L., & Beckman, R. J. (1980). A new family of distributions with applications to Monte Carlo studies. Journal of the American Statistical Association, 75, 276–279.

Kaplansky, I. (1945). A common error concerning kurtosis. Journal of the American Statistical Association, 40, 259.

Lorr, M. (1983). Cluster analysis for social scientists. San Francisco: Jossey Bass Publishers.

Marsaglia, G., Marshall, A. W., & Proshau, F. (1965). Moment crossings as related to density crossings. Journal of the Royal Statistical Society, Series B, 27, 91–93.

McLachlan, G. J. & Basford, K. E. (1988). Mixture models: Inference and applications to clustering. New York: Marcel Dekker.

Meehl, P. E. (1979). A funny thing happened to us on the way to the latent entities. Journal of Personality Assessment, 43, 564–577.

Milligan, G. W. (1979). Ultrametric hierarchical clustering algorithms. Psychometrika, 44, 343-346.

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika, 45, 325–342.

Milligan, G. W. (1985). An algorithm for generating artificial test clusters. Psychometrika, 50, 123-127.

Milligan, G. W. (1989). A validation study of variable weighting algorithm for cluster analysis. Journal of Classification, 6, 53-71.

Milligan, G. W. & Cooper, M. C. (1986). A study of comparability of external criteria for hierarchical cluster analysis. Multivariate Behavioral Research, 21, 441-458.

Milligan, G. W. & Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. Journal of Classification, 5, 181-204.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 66, 846-850.

Rohlf, F. J. (1970). Adaptive hierarchical clustering schemes. Systematic Zoology, 19, 58–82.

SAS (1985). SAS user's guide: Statistics, version 5 edition. Cary, NC: SAS Institute, Inc.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. Journal of the American Statistical Association, 81, 826-831.

Smith, B. T., Boyle, J. M., Garbow, B. S., Ikebe, Y., Klema, V. C., & Moler, C. B. (1974). Matrix eigensystem routines: EISPACK guide. New York: Springer-Verlag.

Stuart, A., & Ord, J. K. (1987). Kendall's advanced theory of statistics: Vol 1. Distributional theory (5th ed.). London: Charles W. Griffin & Co.

Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. New York: Wiley.

Table 1

Effect of Mean Differences and Within–group Variance on

the Kurtosis of Mixtures of Normal Distributions

	$\mu_j \equiv \mu$	$\mu_j \neq \mu$
$\sigma_j^2 \equiv \sigma^2$	(I) $g_2 = 0$	(II) $g_2 < 0$ (unless π_j very different)
$\sigma_j^2 \neq \sigma^2$	(III) $g_2 > 0$	(IV) not determined

Table 2
Screening Measures for Selected Two and Three Subgroup Mixtures

Two Subgroup (σ_1^2, σ_2^2, π_1, π_2)[a]	g_2	m	b
1, 1, .1, .9	.051	.044	.330
1, 1, .3, .7	-.111	-.132	.353
1, 1, .5, .5	-.500	-.500	.400
1, 1, .7, .3	-.607	-.868	.527
1, 1, .9, .1	4.140	-1.044	.866
1, 4, .1, .9	.061	.050	.268
1, 4, .5, .5	.272	.085	.435
1, 4, .9, .1	5.417	-.725	.843
Three Subgroup (σ_1^2, σ_2^2, σ_3^2, π_1, π_2, π_3)[b]			
1, 1, 1, .3 .4 .3	-.664	-.664	.428
1, 1, 1, .1 .6 .3	.343	.203	.341
1, 1, 1, .3 .6 .1	1.941	.051	.585
1, 1, 1, .1 .1 .3	.474	.270	.346
1, 1, 1, .1 .8 .1	.395	.395	.295
1, 1, 4, .3 .4 .3	.359	-.008	.407
1, 1, 4, .1 .6 .3	.280	.378	.296
1, 1, 4, .3 .6 .1	3.529	.663	.579
1, 1, 4, .1 .1 .8	-.018	-.036	.341
1, 1, 4, .1 .8 .1	1.756	1.516	.261
1, 4, 1, .3 .4 .3	-.873	-.873	.470
1, 4, 1, .1 .6 .3	.380	.378	.296
1, 4, 1, .3 .6 .1	.345	-.495	.550
1, 4, 1, .1 .1 .8	.854	.738	.289
1, 4, 1, .1 .8 .1	-.336	-.336	.375

[a] $\mu_1 = 1$, μ_2 is determined such that the overall mean = 0; [b] $\mu_1 = 1$, $\mu_2 = 0$, μ_3 is determined such that the overall mean = 0.

Table 3

Expected Values of Screening Measures for Selected Distributions

	g_1	g_2	m	b
Bernoulli (p)	$\frac{1 - 2p}{\sqrt{p(1 - p)}}$	$\frac{1}{p(1 - p)} - 6$	-2	1
Binomial (n,p)	$\frac{1 - 2p}{\sqrt{np(1 - p)}}$	$\frac{1 - 6p + 6p^2}{np(1 - p)}$	$-\frac{2}{n}$	$\frac{1 - 4p + 4p^2 + np - np^2}{1 - 6p + 6p^2 + 3np - 3np^2}$
Geometric (p)	$\frac{2 - p}{\sqrt{1 - p}}$	$6 + \frac{p^2}{1-p}$	2	$\frac{5 - 5p + p^2}{9 - 9p + p^2}$
Poisson (m)	$\frac{\sqrt{m}}{m}$	$\frac{1}{m}$	0	$\frac{m + 1}{3m + 1}$
Exponential (λ)	2	6	2	.556
Normal (μ,σ^2)	0	0	0	.333
χ^2 (k)	$\frac{2\sqrt{2k}}{k}$	$\frac{12}{k}$	$\frac{4}{k}$	$\frac{k + 8}{3k + 12}$
t (n)	0	$\frac{6}{n - 4}$	$\frac{6}{n - 4}$	$\frac{n - 4}{3n - 6}$
Uniform (a,b)	0	-1.2	-1.2	.556

Table 4

Relative Effectiveness of Variable Selection/Weighting Methods

Method	Within-solution Ordinal Change			Ordinal Comparison		Paired t-test	
	d_w	z	p	z	p	t(1727)	p
$g_2 < -1.2$	-.273	-17.29	< .001	-17.39	< .001	-17.32	< .001
$g_2 < 0$, signif.	-.108	-7.06	< .001	-6.88	< .001	-10.19	< .001
$g_2 < 0$	-.094	-7.07	< .001	-7.58	< .001	-9.16	< .001
$m < -1.2$	.047	3.86	< .001	5.47	< .001	4.91	< .001
$m < 0$	.050	5.18	< .001	6.10	< .001	5.38	< .001
$b > .555$	.072	5.75	< .001	6.69	< .001	5.82	< .001
$b > .333$	.030	4.12	< .001	5.10	< .001	4.63	< .001
Ultrametric weights	.086	7.09	< .001	7.43	< .001	6.95	< .001
Forward Selection	.010	0.76	> .05	0.93	> .05	5.39	< .001

Table 5

Mean HA-Rand Recovery Values for
Variable Selection/Weighting Methods by Number of Error Dimensions

Method	Number of Error Dimensions				Overall Mean	Rank (Ordinal Method)	Rank (Paired t-tests)
	0	1	2	3			
$b > .555$	.942	.942	.939	.939	.941	1^{a}	1^{f}
Ultrametric Weights	.962	.937	.938	.926	.941	2^{ab}	1^{f}
$m < -1.2$	.948	.937	.930	.920	.933	3^{bc}	4^{fg}
$m < 0$	.983	.938	.911	.880	.928	4^{c}	4^{fg}
$b > .333$	.986	.934	.892	.870	.920	5^{d}	6^{g}
Forward Selection	.966	.927	.925	.920	.934	6^{de}	1^{f}
No Selection	.984	.917	.877	.852	.908	7^{de}	7

^{abcdefg} Methods with common superscripts do not significantly differ from one another. Ranks are based on the number of methods that were significantly lower.

Table 6

Means and (Std. Dev.) of HA-Rand Recovery Values for
Variable Selection/Weighting Methods by Type of Error Condition

Method	Low Error	Ranks	High Error	Ranks	20% Outlier	Ranks
No Selection	.934 (.216)	1[a]	.838 (.304)	1[a]	.972 (.146)	1
$b > .333$	.936 (.214)	1[a]	.841 (.294)	1[a]	.959 (.178)	2[g]
$m < 0$	.923 (.237)	1[ab]	.819 (.313)	3[f]	.962 (.151)	2[g]
Ultrametric Weights	.922 (.209)	4[bc]	.827 (.285)	3[f]	.956 (.158)	2[g]
Forward Selection	.941 (.159)	4[c]	.849 (.226)	3[f]	.941 (.172)	5
$b > .555$	.899 (.219)	6[d]	.750 (.302)	6	.873 (.256)	6[h]
$m < -1.2$	.886 (.244)	6[d]	.721 (.298)	7	.894 (.204)	7[h]

[abcdefgh] Methods with common superscripts do not significantly differ from one another. Ranks are based on the number of methods that were significantly lower.

Table 7

Effectiveness of Combination of Ultrametric Weights and Screening Methods

Combination Method	Comparison	Within-solution Ordinal Change			Ordinal Comparison of Distributions		Paired t-test	
		d_w	z	p	z	p	t(1727)	p
$b > .555$, weights	Screening only	.024	2.18	.029	1.36	> .05	5.03	< .001
	Ultrametric Weights only	.030	2.71	.007	2.03	.042	4.61	< .001
$m < -1.2$, weights	Screening only	-.008	-0.62	> .05	-2.05	.040	3.19	< .001
	Ultrametric Weights only	-.028	-2.48	.013	-3.07	.002	1.67	> .05
$b > .333$, weights	Screening only	.071	6.13	< .001	5.86	< .001	5.93	< .001
	Ultrametric Weights only	.017	3.61	< .001	3.61	< .001	3.10	.002
$m < 0$, weights	Screening only	.067	5.84	< .001	5.30	< .001	5.67	< .001
	Ultrametric Weights only	.035	5.62	< .001	5.30	< .001	4.80	< .001

Table 8

Results for Combined Weights and Screening Methods

Method	Number of Error Dimensions				Overall Mean	Rank (Ordinal Method)	Rank (Paired t-tests)
	0	1	2	3			
$m < 0$, weights	.969	.963	.949	.937	.954	1^{a}	1^{g}
$b > .555$, weights	.962	.962	.961	.962	.962	2^{abc}	1^{g}
$b > .555$	.942	.942	.939	.939	.941	3^{bc}	4^{hi}
$b > .333$, weights	.967	.954	.940	.929	.948	4^{bd}	3^{h}
Ultrametric Weights	.962	.937	.938	.926	.941	4^{ce}	5^{i}
$m < -1.2$	.948	.937	.930	.920	.933	6^{de}	5^{i}
Forward Selection	.966	.927	.925	.920	.934	7^{f}	5^{i}
No Selection	.984	.917	.877	.852	.908	7^{f}	8

^{abcdefghi} Methods with common superscripts do not significantly differ from one another. Ranks are based on the number of methods that were significantly lower.

Table 9
Means and (Std. Dev.) of HA-Rand Recovery Values for
Combined Variable Selection/Weighting Methods by Type of Error Condition

Method	Low Error			High Error			20% Outlier		
	Mean (Std. Dev.)	Ordinal Ranks	t-Test Ranks	Mean (Std. Dev.)	Ordinal Ranks	t-Test Ranks	Mean (Std. Dev.)	Ordinal Ranks	t-Test Ranks
No Selection	.934 (.216)	1	1^{e}	.838 (.304)	1	1^{h}	.973 (.146)	1	1^{m}
Ultrametric Weights	.922 (.209)	2^{a}	3^{cd}	.827 (.285)	2^{f}	1^{h}	.956 (.158)	2	2^{mn}
$m < 0$, weights	.922 (.202)	2^{a}	3^{cd}	.820 (.283)	2^{f}	1^{h}	.947 (.154)	3^{j}	3^{mno}
$b > .333$, weights	.927 (.193)	2^{a}	3^{cd}	.824 (.284)	2^{f}	1^{h}	.935 (.199)	3^{j}	5^{o}
Forward Selection	.941 (.159)	2^{a}	1^{e}	.849 (.226)	2^{f}	1^{h}	.941 (.172)	4^{jk}	4^{no}
$b > .555$	.899 (.219)	6	6^{de}	.750 (.302)	6	6^{i}	.873 (.256)	7^{l}	7^{pq}
$m < -1.2$	.886 (.244)	7^{b}	7^{e}	.721 (.298)	7^{g}	6^{i}	.894 (.204)	7^{l}	8^{q}
$b > .555$, weights	.872 (.240)	7^{b}	7^{e}	.716 (.293)	7^{g}	6^{i}	.924 (.188)	6^{k}	6^{op}

^{abcdefghijklmnop} Methods with common superscripts do not significantly differ from one another. Ranks are based on the number of methods that were significantly lower.

Table 10

Comparison of Forward Selection Using Unstandardized Variables with Other Methods

Comparison	Within-solution Ordinal Change			Ordinal Comparison of Distributions		Paired t-test	
	d_w	z	p	z	p	t(1727)	p
Forward Selection, standardized	.086	8.33	< .001	9.55	< .001	6.47	< .001
b > .555, weighted	-.013	-1.18	> .05	-0.67	> .05	-2.32	< .021
m < 0, weighted	-.013	-1.22	> .05	-0.99	> .05	-0.34	> .05
Ultrametric weights	.013	1.20	> .05	1.23	> .05	2.52	< .012
No Selection	.089	9.68	< .001	10.26	< .001	10.21	< .001

Table 11

Comparison of Forward Selection Using Unstandardized Variables with Other Methods

Method	Number of Error Dimensions				Overall Mean	Ordinal Ranks	Paired t-test Ranks
	0	1	2	3			
$m < 0$, weights	.969	.963	.949	.937	.954	1^{a}	1^{g}
$b > .555$, weights	.962	.962	.961	.962	.962	2^{abc}	1^{g}
$b > .555$	.942	.942	.939	.939	.941	2^{abc}	5^{hi}
Forward Selection, Unstandardized	.986	.948	.942	.936	.953	4^{bc}	3^{ghi}
$b > .333$, weights	.967	.954	.940	.929	.948	4^{bd}	4^{h}
Ultrametric Weights	.962	.937	.938	.926	.941	6^{ce}	6^{ij}
$m < -1.2$	.948	.937	.930	.920	.933	7^{bde}	7^{j}
Forward Selection, Standardized	.966	.927	.925	.920	.934	8^{f}	7^{j}
No Selection	.984	.917	.877	.852	.908	8^{f}	9

^{abcdefghij} Methods with common superscripts do not significantly differ from one another. Ranks are based on the number of methods that were significantly lower.

Appendix

Let X be distributed as a finite mixture composed of G distributions with mixing proportions π_j ($j=1,G$), i.e.:

$$f(X) = \sum_{j=1}^{G} \pi_j f_j(X) \quad .$$

We will be particularly interested in the normal mixture model, in which each of the subgroup distributions $f_j(X)$ is $\sim N(\mu_j,\sigma_j^2)$. Let μ be the grand mean of X (i.e. $\mu = \Sigma\pi_j\mu_j$). Let M_k be the kth central moment for the mixture, and M_{kj} is the k^{th} central moment for subgroup j: $M_{kj} = \mathscr{E}(X - \mu_j)^k$. Finally, by the linearity of the expectation operator, note that:

$$\mathscr{E}[f(X)] = \sum_{j=1}^{G} \pi_j \mathscr{E}_j[f(X)] \quad .$$

where $\mathscr{E}_j$ is the expectation with respect to the distribution of subgroup j.

Kurtosis

For a mixture, the fourth moment of the mixture distribution is:

$$M_4 = \sum_{j=1}^{G} \pi_j \mathscr{E}_j(X - \mu)^4$$

$$M_4 = \sum_{j=1}^{G} \pi_j \mathscr{E}_j\left((X - \mu_j) + (\mu_j - \mu)\right)^4$$

Expanding the binomial and taking expectations yields:

$$M_4 = \sum_{j=1}^{G} \pi_j M_{4j} + 4\sum_{j=1}^{G} \pi_j M_{3j}(\mu_j - \mu) + 6\sum_{j=1}^{G} \pi_j \sigma_j^2(\mu_j - \mu)^2 + \sum_{j=1}^{G} \pi_j(\mu_j - \mu)^4 \quad .$$

Similarly, the second moment is:

$$M_2 = \sum_{j=1}^{G} \pi_j \sigma_j^2 + \sum_{j=1}^{G} \pi_j(\mu_j - \mu)^2 \quad . \qquad \text{(A)}$$

Using (1), the kurtosis of the mixture is:

$$g_2 = \frac{\sum_{j=1}^{G} \pi_j M_{4j} + 4\sum_{j=1}^{G} \pi_j M_{3j}(\mu_j - \mu) + 6\sum_{j=1}^{G} \pi_j \sigma_j^2 (\mu_j - \mu)^2 + \sum_{j=1}^{G} \pi_j (\mu_j - \mu)^4}{\left[\sum_{j=1}^{G} \pi_j \sigma_j^2 + \sum_{j=1}^{G} \pi_j (\mu_j - \mu)^2\right]^2} - 3 \ .$$

Let $k = g_2 M_2^2$. Then expanding and collecting like terms yields:

$$k = \sum_{j=1}^{G} \pi_j M_{4j} - 3\left(\sum_{j=1}^{G} \pi_j \sigma_j^2\right)^2$$

$$+ 4\sum_{j=1}^{G} \pi_j M_{3j}(\mu_j - \mu)$$

$$+ 6\sum_{j=1}^{G} \pi_j \sigma_j^2 (\mu_j - \mu)^2 - 6\left(\sum_{j=1}^{G} \pi_j \sigma_j^2\right)\left(\sum_{j=1}^{G} \pi_j (\mu_j - \mu)^2\right)$$

$$+ \sum_{j=1}^{G} \pi_j (\mu_j - \mu)^4 - 3\left(\sum_{j=1}^{G} \pi_j (\mu_j - \mu)^2\right)^2 \ .$$

$$k = \sum_{j=1}^{g} \pi_j M_{4j} - 3\sum_{j=1}^{G} \pi_j \sigma_j^4 + 4\sum_{j=1}^{G} \pi_j M_{3j}(\mu_j - \mu) - 2\sum_{j=1}^{G} \pi_j (\mu_j - \mu)^4$$

$$+ 3var[\sigma_j^2] + 6cov[\sigma_j^2, (\mu_j - \mu)^2] + 3var[(\mu_j - \mu)^2] \qquad \text{(B)}$$

$$= \sum_{j=1}^{g} \pi_j M_{4j} - 3\sum_{j=1}^{G} \pi_j \sigma_j^4 + 4\sum_{j=1}^{G} \pi_j M_{3j}(\mu_j - \mu) + 3var[\sigma_j^2 + (\mu_j - \mu)^2] - 2\sum_{j=1}^{G} \pi_j (\mu_j - \mu)^4$$

The above expression (B) is valid for any distribution which has the first four moments. Making explicit use of the fact that the subgroups are normal, $M_{4j} = 3\sigma_j^4$ and $M_{3j} = 0$ for all j. Making these substitutions:

$$k = 3var[\sigma_j^2 + (\mu_j - \mu)^2] - 2\sum_{j=1}^{G} \pi_j (\mu_j - \mu)^4$$

and hence:

$$g_2 = \frac{3\,var\,[\sigma_j^2 + (\mu_j - \mu)^2] - 2\sum_{j=1}^{G} \pi_j(\mu_j - \mu)^4}{\left[\sum_{j=1}^{G} \pi_j \sigma_j^2 + \sum_{j=1}^{G} \pi_j(\mu_j - \mu)^2\right]^2} \qquad \text{(C)}$$

Skewness

The skewness of a variable is:

$$g_1 = \frac{\sum_{j=1}^{G} \pi_j \mathcal{E}(X-\mu)^3}{[M_2]^{\frac{3}{2}}}$$

$$= \frac{\sum_{j=1}^{G} \pi_j \mathcal{E}[(X-\mu_j) + (\mu_j - \mu)]^3}{\left[\sum_{j=1}^{G} \pi_j \sigma_j^2 + \sum_{j=1}^{G} \pi_j(\mu_j - \mu)^2\right]^{\frac{3}{2}}}$$

$$g_1 = \frac{\sum_{j=1}^{G} \pi_j M_{3j} + 3\sum_{j=1}^{G} \pi_j \sigma_j^2(\mu_j - \mu) + \sum_{j=1}^{G} \pi_j(\mu_j - \mu)^3}{\left[\sum_{j=1}^{G} \pi_j \sigma_j^2 + \sum_{j=1}^{G} \pi_j(\mu_j - \mu)^2\right]^{\frac{3}{2}}} \quad . \qquad \text{(D)}$$

Assuming within–group normality, this reduces to:

$$g_1 = \frac{3\sum_{j=1}^{G} \pi_j \sigma_j^2(\mu_j - \mu) + \sum_{j=1}^{G} \pi_j(\mu_j - \mu)^3}{\left[\sum_{j=1}^{G} \pi_j \sigma_j^2 + \sum_{j=1}^{G} \pi_j(\mu_j - \mu)^2\right]^{\frac{3}{2}}} \quad . \qquad \text{(E)}$$

Table A1

Main Effects for Design Factors in the Simulation Study

Number of Subgroups	Mean	Std.
2	.848	.348
3	.925	.189
4	.915	.197
5	.876	.219

Number of Core Variables	Mean	Std.
4	.810	.295
6	.914	.228
8	.950	.190

Number of Error Dimensions	Mean	Std.
0	.928	.217
1	.894	.244
2	.879	.259
3	.864	.268

Subgroup Sizes	Mean	Std.
Equal	.949	.153
60% in one subgroup	.862	.302
10% in one subgroup	.862	.257

Clustering Algorithm	Mean	Std.
Average Linkage	.912	.222
Ward's Method	.906	.244
Complete Linkage	.876	.253
Single Linkage	.871	.270

Table A1 (cont.)

Weighting/Selection	Mean	Std.
No Selection	.908	.214
Ultrametric Weights	.941	.178
Forward Selection	.934	.149
m < 0	.930	.191
m < -1.2	.933	.187
b > .333	.920	.200
b > .555	.941	.171
g_2 < 0	.839	.311
g_2 signif. < 0	.816	.343
g_2 < -1.2	.750	.354

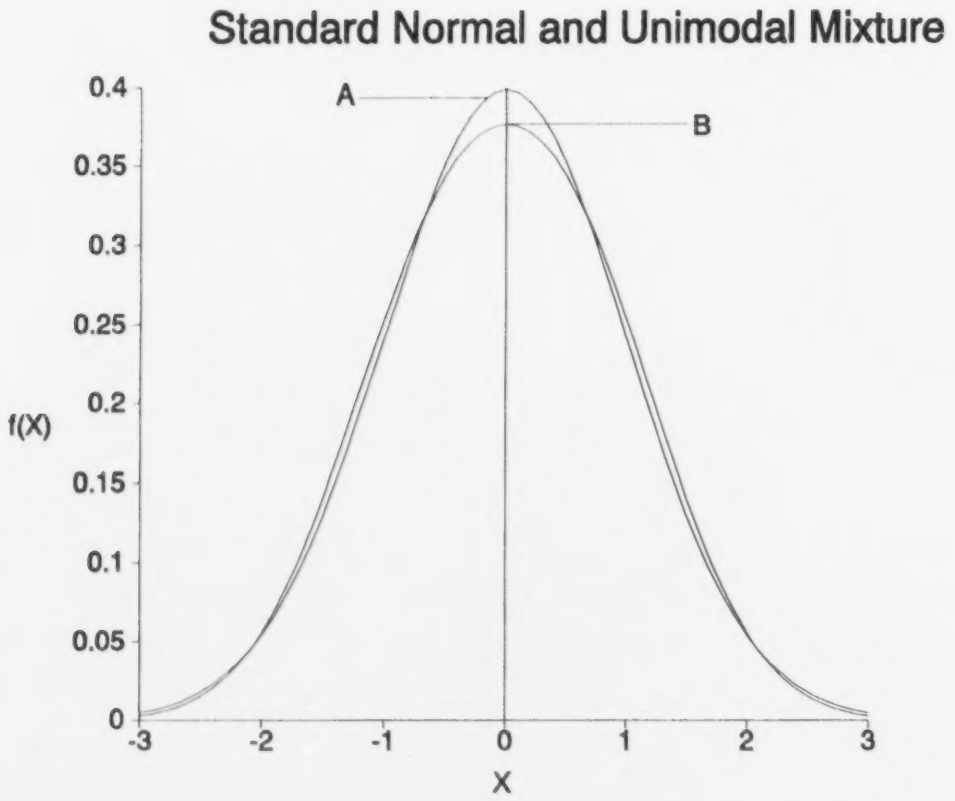


Figure 1. Plot of A) standard normal (0,1) density and B) unimodal mixture of two normal densities with same overall mean and variance:

$$f_A(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}, \quad f_B(x) = \frac{1}{\sqrt{1.28\pi}} \left[0.5e^{\frac{-(x+0.6)^2}{1.28}} + 0.5e^{\frac{-(x-0.6)^2}{1.28}} \right].$$

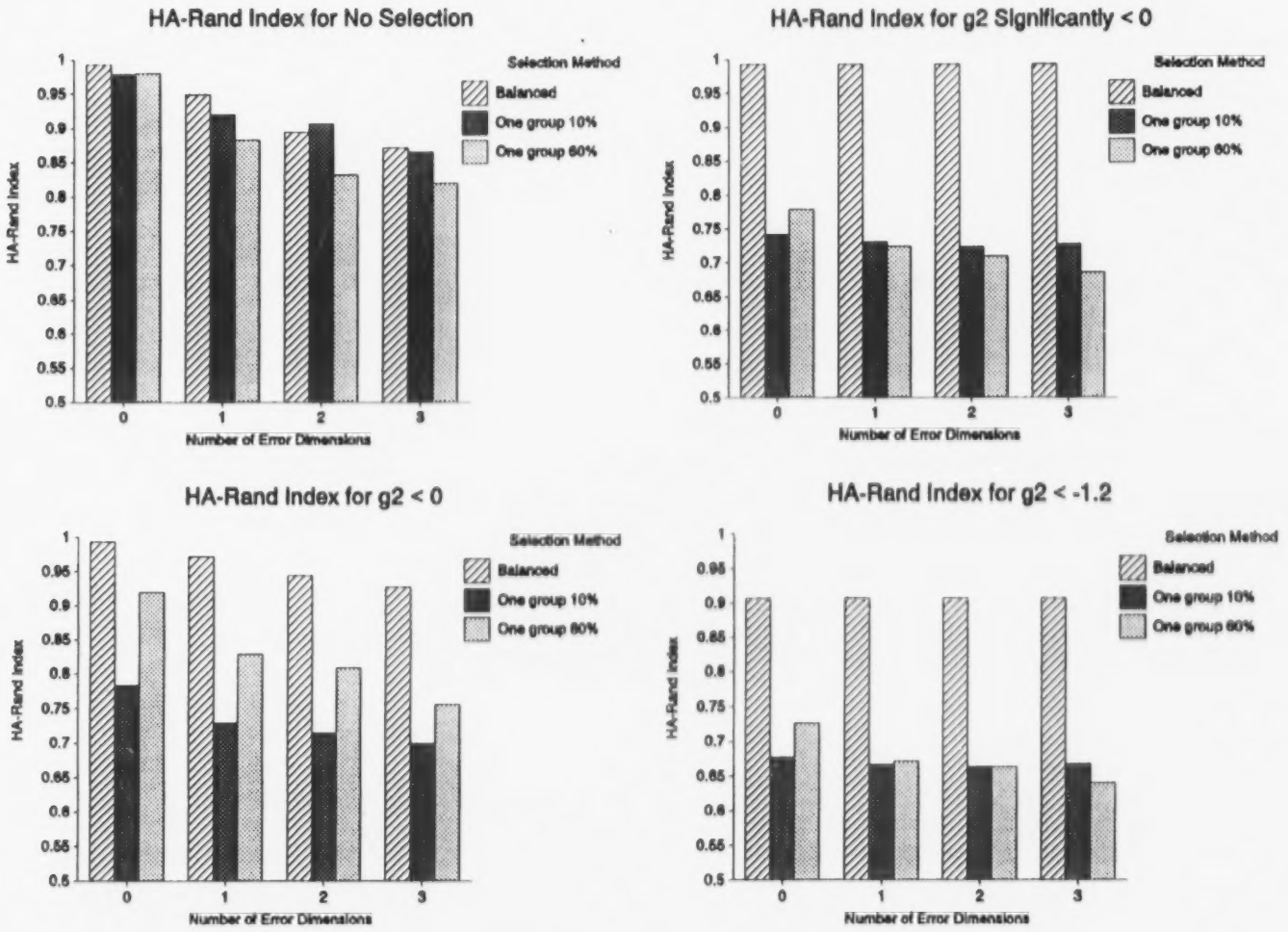
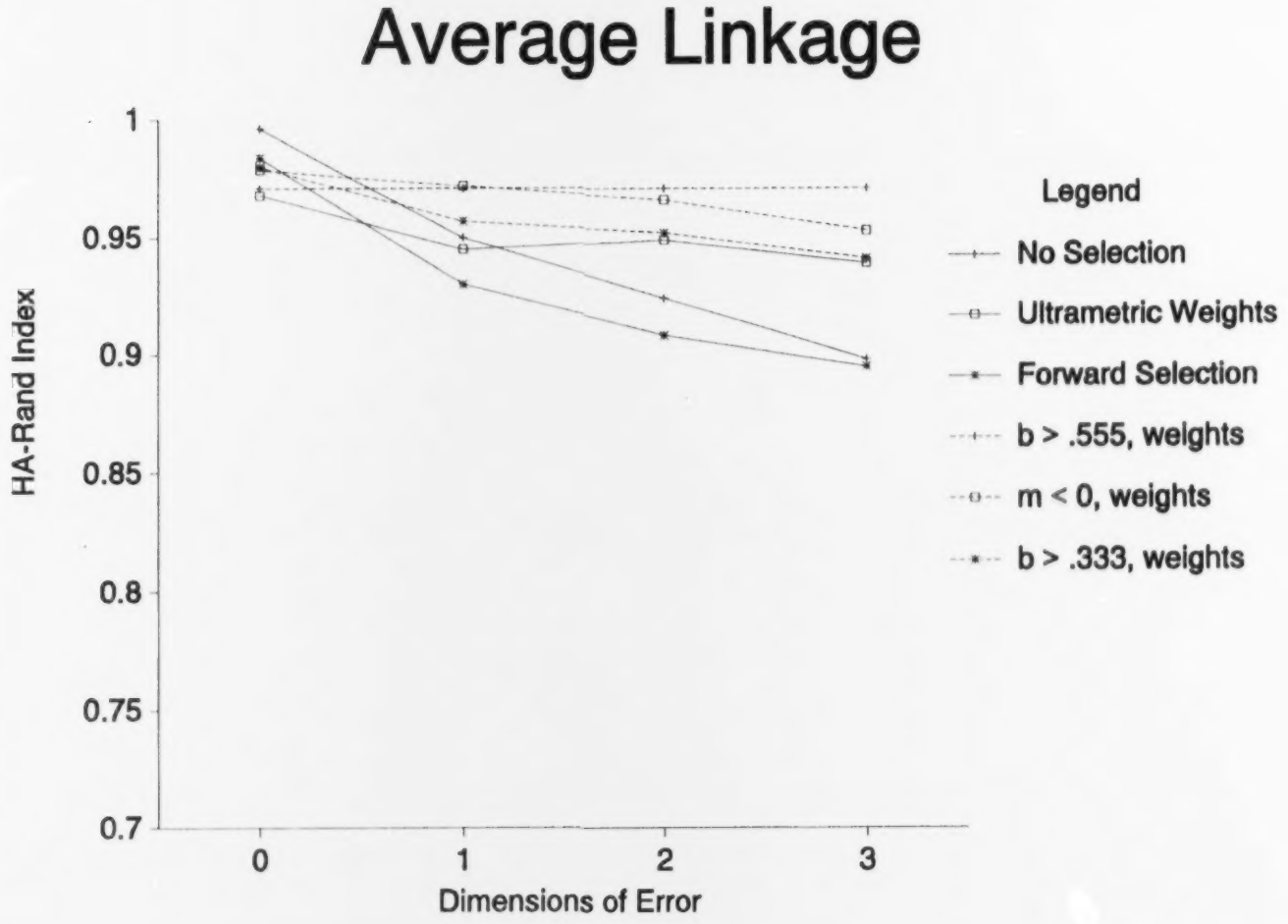


Figure 2. Mean HA-Rand Index for No Selection and Kurtosis-based screening by subgroup size.



61

65

Figure 3. Interaction of number of irrelevant dimensions by variable weighting/selection method: Average Linkage.

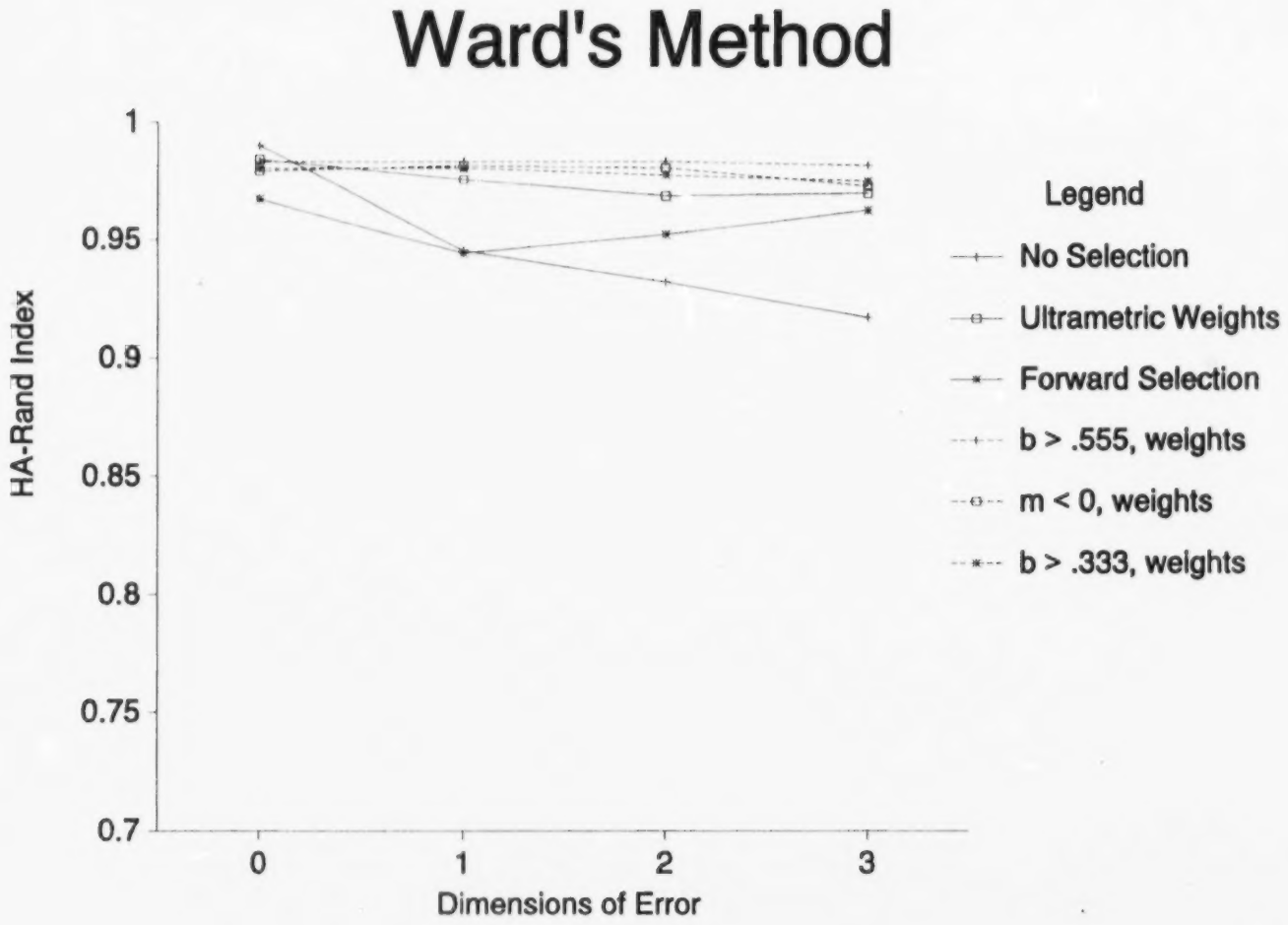


Figure 4. Interaction of number of irrelevant dimensions by variable weighting/selection method: Ward's Method.

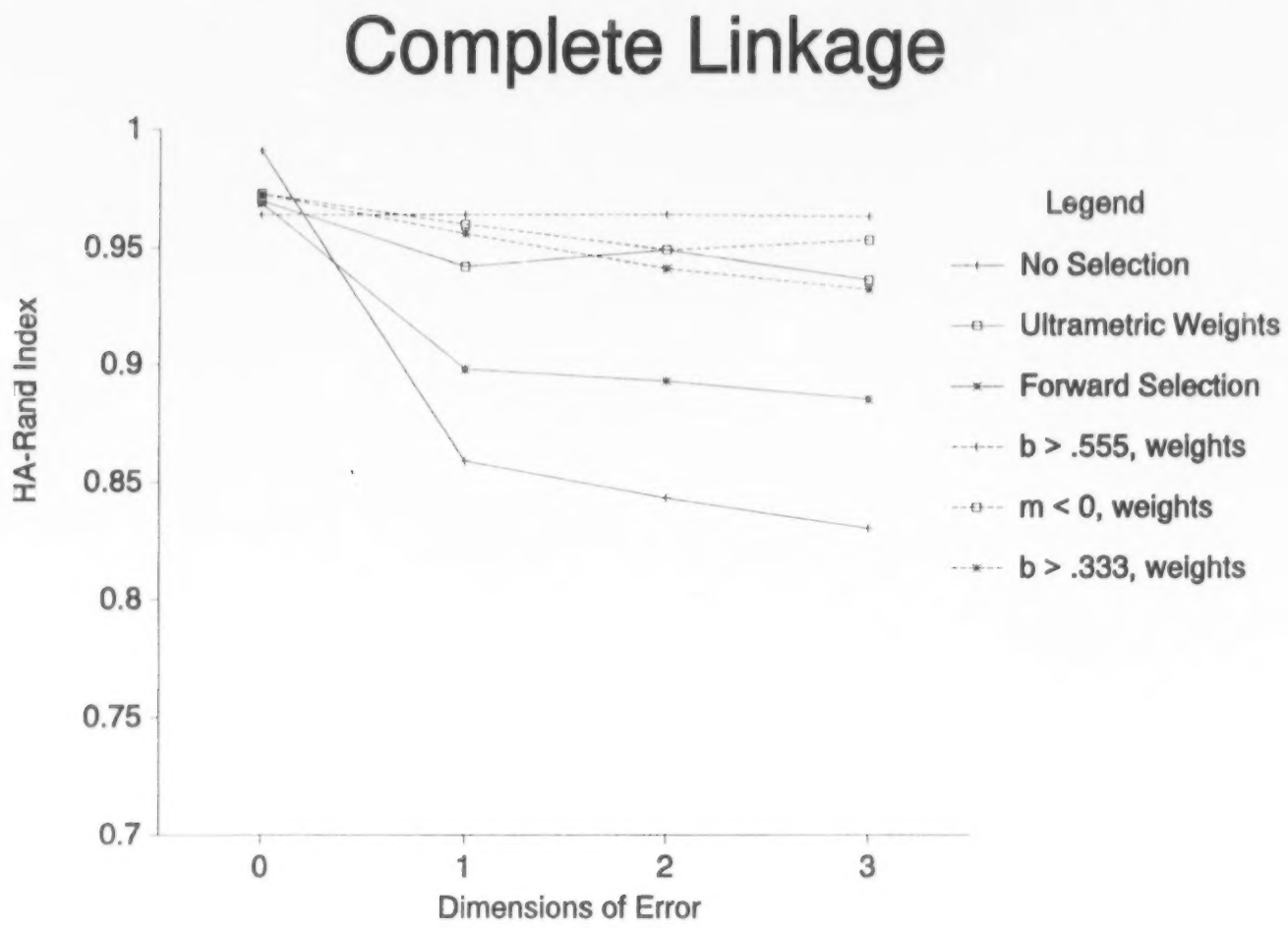


Figure 5. Interaction of number of irrelevant dimensions by variable weighting/selection method: Complete Linkage.

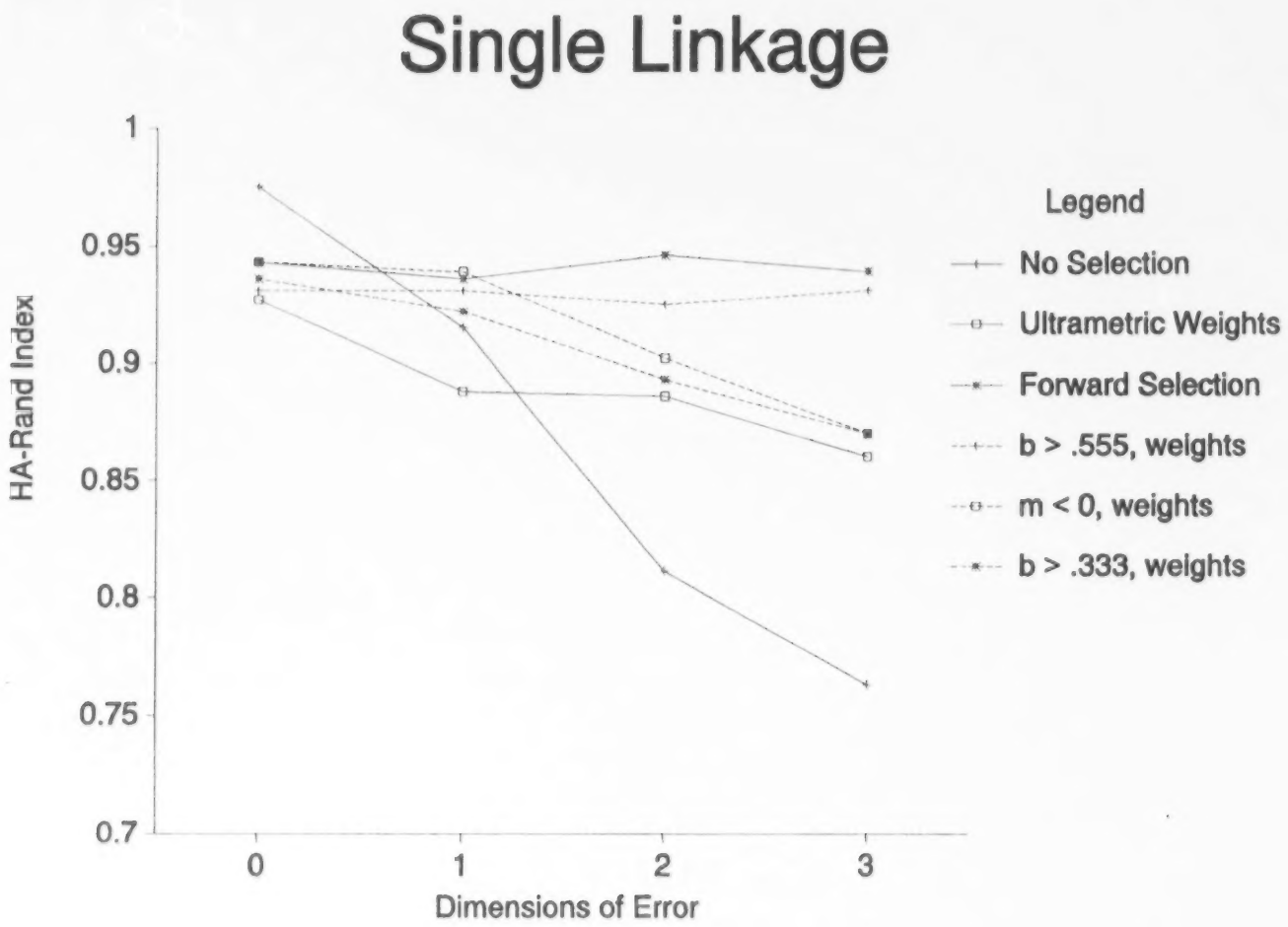


Figure 6. Interaction of number of irrelevant dimensions by variable weighting/selection method: Single Linkage.